R²F: A General Retrieval, Reading and Fusion Framework for Document-level Natural Language Inference

Hao Wang[§], **Yixin Cao**[†*], **Yangguang Li**[‡], **Zhen Huang**[§]
Kun Wang[‡], **Jing Shao**[‡]
[§] National University of Defense Technology
[†] Singapore Management University [‡] SenseTime

Abstract

Document-level natural language inference (DOCNLI) is a new challenging task in natural language processing, aiming at judging the entailment relationship between a pair of hypothesis and premise documents. Current datasets and baselines largely follow sentence-level settings, but fail to address the issues raised by longer documents. In this paper, we establish a general solution, named Retrieval, Reading and Fusion (R²F) framework, and a new setting, by analyzing the main challenges of DOCNLI: interpretability, long-range dependency, and cross-sentence inference. The basic idea of the framework is to simplify document-level task into a set of sentence-level tasks, and improve both performance and interpretability with the power of evidence. For each hypothesis sentence, the framework retrieves evidence sentences from the premise, and reads to estimate its credibility. Then the sentence-level results are fused to judge the relationship between the documents. For the setting, we contribute complementary evidence and entailment label annotation on hypothesis sentences, for interpretability study. Our experimental results show that R²F framework can obtain state-of-the-art performance and is robust for diverse evidence retrieval methods. Moreover, it can give more interpretable prediction results. Our model and code are released at https://github.com/phoenixsecularbird/R2F.

1 Introduction

Natural Language Inference (NLI) is the task of determining whether a *hypothesis* is entailed or not in a *premise*. While earlier works (Bowman et al., 2015; Williams et al., 2018; Wang et al., 2019; Nie et al., 2020) assume that both hypothesis and premise are single sentences, recent research pays more attention on document-level task, namely Document-level NLI (DOCNLI) (Yin et al., 2021). The task can enlarge the task scope to judge the

variability of semantic expression for many Natural Language Processing (NLP) tasks, e.g., exposure bias (Bengio et al., 2015; Arora et al., 2022) alleviation for text summarization (Sandhaus, 2008; Narayan et al., 2018), and human-manipulated news articles recognition for automatic fake news detection (Jawahar et al., 2022; Huang et al., 2022).

Compared with sentence-level NLI, DOCNLI poses many new challenges, while there are only a few datasets and models. In terms of datasets, Yin et al. (2021) reformat some mainstream NLP tasks, e.g., text summarization and question answering, and build the *first* large scale dataset DOCNLI with over 1 million document pairs[1]. However, the dataset does not provide evidence annotation about how the labels are inferred, i.e., which hypothesis sentences lead to semantic inconsistency, or which premise sentences help to decide the entailment relationship. As shown in Figure 1, although the sample is annotated as *not entailment*, most of the hypothesis sentences are actually entailed. By contrast, the detailed disinformation of "*in 1989*" in the third hypothesis sentence eventually decides the entailment relationship between the documents. For each hypothesis sentence, only several premise sentences are enough to serve as the exact evidence to judge its own sentence-level entailment label.

In this paper, we argue that evidence discovery is important and challenging for DOCNLI. Our pilot experiments in Section 4.3 and 4.5 show that randomly selected evidences can still contribute to comparable performance. Thus, only the black-box models may be not so convincing. However, to annotate evidence for evaluation is non-trivial. For each hypothesis sentence, on one hand, it may refer to multiple premise sentences. On the other hand, there may be several evidence groups, where each group can independently serve the label prediction. We highlight this as **interpretability** challenge.

---

[1] Yin et al. (2021) propose the task and annotate the dataset with the same name DocNLI.

*Corresponding Author.

Label: not_entailment

Hypothesis: ①US cities along the Gulf of Mexico from Alabama to eastern Texas were on alert last night as Hurricane Andrew headed west after hitting southern Florida leaving at least eight dead, causing severe property damage, and leaving 1.2 million homes without electricity. ②Gusts of up to 165 mph were recorded. ③It is the fiercest hurricane to hit the US **in 1989**. ④As Andrew moved across the Gulf there was concern that it might hit New Orleans, which would be particularly susceptible to flooding, or smash into the concentrated offshore oil facilities. ⑤President Bush authorized federal disaster assistance for the affected areas.

Premise: ①US CITIES along the Gulf of Mexico from Alabama to eastern Texas were on storm watch last night as Hurricane Andrew headed west after sweeping across southern Florida, causing at least eight deaths and severe property damage. ③The hurricane was one of the fiercest in the US **in decades** and the first to hit Miami directly in a quarter of a century. ⋯ ④However, Hurricane Andrew gathered fresh strength as it moved across the Gulf of Mexico and there was concern last night that it might head towards New Orleans, which is especially low lying and could suffer severe flood damage. ⋯ ④It could threaten the large concentration of offshore oil production facilities in the Gulf of Mexico. ⋯ ②Andrew, the first Caribbean hurricane of the season, hit the eastern coast of Florida early yesterday, gusting up to 165mph. ⋯ ①The Florida Power and Light company said that about 1.2m of its customers, or 32 per cent, were without power. ⋯ ⑤President Bush authorized federal disaster assistance for the affected areas and made plans for an inspection tour of the state. ⋯

Figure 1: A sample of DOCNLI dataset. For each sample, only the entailment label between the documents is annotated. For display, we mark each hypothesis sentence and its corresponding premise sentences (namely the evidences, not annotated in the original dataset) with the same number and color. The sample is annotated as *not entailment* due to the disinformation of "*in 1989*" in the third hypothesis sentence. The premise is partly omitted.

In term of models, current baselines (Yin et al., 2021; Zhong et al., 2020) still largely follow sentence-level NLI. They either concatenate two documents for mutual information interaction for classification, or encode them separately for semantic match with document-level representations. However, except for the interpretability issue, they will leave the following challenges unexplored:

• **Long-range Dependency Modeling** The task requests to handle a pair of long documents at the same time, where we observe that 29.81% samples of DOCNLI dataset (Yin et al., 2021) contain more than 500 *words*[2], while 10.47% samples contain more than 1000 *words*. This will not only exceed the input limit of Pre-trained Language Models (PLMs), but also make it far more difficult to capture long-range dependency. Necessary information interaction between the hypothesis and some key premise sentences may not be guaranteed. Besides, most contexts are uninformative for entailment inference and will only serve as noise.

• **Cross-sentence Inference** To judge the relationship between the documents, it is supposed to consider all hypothesis sentences, where the detailed disinformation issue still remains unsolved. Besides, the verification of one hypothesis sentence may request to combine multiple and distant premise sentences, different from the sentence pair mode in sentence-level NLI. As shown in Figure 1,

to process the first one, it needs to take both the first and sixth premise sentences (all in red fonts).

In this paper, we establish a general solution, named Retrieval, Reading and Fusion (R^2F) framework, and a new setting for the task. The basic idea of the framework is to simplify document-level classification task into a set of sentence-level tasks, and then improve both performance and interpretability with the power of evidence. Specifically, the framework splits the hypothesis document into sentences. Then for each hypothesis sentence, it retrieves evidence sentences from the premise, and reads to estimate its credibility score upon the evidences. Finally, it fuses the sentence-level results and judge the entailment relationship between the two documents. For the setting, we contribute complementary fine-grained annotations for interpretability study. For each hypothesis sentence, we manually annotate entailment label and several evidence groups, where each group is enough to independently infer the label. In summary, our contributions are as follows:

• We propose a Retrieval, Read and Fusion framework as a general solution for DOCNLI task.

• We contribute complementary evidence and entailment label annotation for each hypothesis sentence on a subset of DOCNLI dataset for interpretability study.

• Our experimental results on DOCNLI dataset indicate that the framework obtains state-of-the-

---

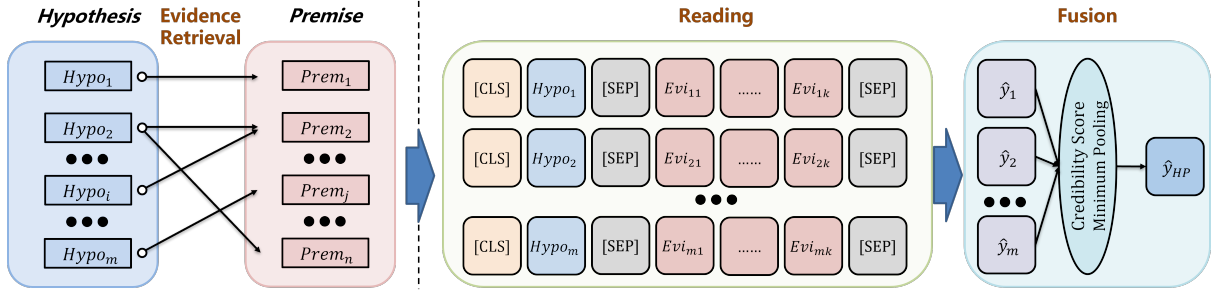[2]A word may correspond to multiple tokens for PLMs.

Figure 2: R²F framework. For each hypothesis sentence, the framework firstly **retrieves** evidence sentences from the premise, and then **reads** to estimate the credibility upon the evidences. Finally it **fuses** the sentence-level results, and judges the entailment relationship between these two documents. $\hat{y}_i$ and $\hat{y}_{HP}$ are the credibility score of the i-th hypothesis sentence and the sample, while Evi_{ij} is the j-th evidence of the i-th hypothesis sentence.

art performance. Besides, it is robust for diverse retrieval methods. Moreover, the framework can give more interpretable prediction results.

2 R²F Framework

Our R²F framework aims at a general solution for DOCNLI task with interpretability, i.e., to obtain corresponding evidence and predict entailment label for each hypothesis sentence. As shown in Figure 2, the framework consists of 3 components, namely evidence retrieval, reading for credibility estimation, and credibility fusion. For efficiency, the retrieval component is an independent unit to provide evidence input for the other two components, which are optimized jointly.

2.1 Task Formulation

Similar to previous sentence-level NLI tasks, for each sample in DOCNLI task, given a hypothesis document H and a premise document P, it is requested to judge the entailment relationship R between these two documents. Here, $R \in \{$"*entailment*", "*not_entailment*"$\}$ for DOCNLI dataset, but may not be restricted to binary classification.

2.2 Evidence Retrieval

Given each sample, we split the hypothesis into sentences and retrieve evidence sentences from the premise. Formally, we split the hypothesis H and the premise P into single sentences $\{H_1, H_2, \cdots, H_m\}$ and $\{P_1, P_2, \cdots, P_n\}$, through NLTK tool[3]. Here, m and n are the sentence numbers.

For each hypothesis sentence H_i, we respectively utilize the following retrieval methods to calculate the *relevance score* with all premise sentences. Then according to the scores, we remain top

K sentences as the corresponding evidence. The value of K is a trade-off between evidence recall and precision. A lower value pursues higher evidence precision, but may lead to evidence missing, while a higher value guarantees higher evidence recall, but may introduce too many uninformative sentences as noise. Moreover, to keep and utilize contextual information, for each hypothesis sentence, we reorder the evidence sentences according to their original order in the premise.

To calculate the relevance score, we take several sparse and dense retrieval methods into consideration:

• **ROUGE-1** Inspired by Mao et al. (2022) and Zhang et al. (2022), we adopt ROUGE-1 retrieval. For a pair of sentences, this sparse retrieval method focuses on *n-gram match* of the pair to calculate ROUGE-1 score as the relevance metric. We take it as the **main retrieval method**.

• **BM25**[4] BM25 is one of the most advanced sparse retrieval methods. We take all premise sentences as the corpus. For a pair of sentences, BM25 involves not only the pair itself but also the whole corpus, to count *term frequency* and *inverse-document frequency* to obtain the relevance score.

• **SimCSE**[5] Inspired by Gao et al. (2021), we utilize SimCSE (Gao et al., 2021), a strong sentence embedding model, as dense retrieval method for *semantic match*. For a pair of sentences, we take the cosine similarity of the sentence embeddings as the relevance score.

Except for above retrieval methods, we also adopt another simple but effective strategy. If a hypothesis sentence is a *substring* of the premise,

---

[3]https://www.nltk.org/

[4]We adopt the implementation from https://github.com/dorianbrown/rank_bm25.

[5]We adopt unsupervised and supervised version of RoBERTa$_{base}$.

then it is naturally entailed in the premise through *string match* and does not need further study.

2.3 Reading & Credibility Estimation

For each hypothesis sentence, we concentrate informative premise sentences and filter out most noisy ones through evidence retrieval. Then this component aims to estimate its credibility, which involves the hypothesis sentence itself and several evidence sentences. This is different from conventional sentence-level NLI, which studies the relationship between a pair of sentences. Hence we adopt reading models. Our general R^2F framework is compatible to any arbitrary reading models, which may be enhanced by several advanced learning technologies, e.g., graph neural network (Kipf and Welling, 2017; Velickovic et al., 2018), commonsense knowledge injection (Zhang et al., 2019; Wang et al., 2021), and syntactic structure analysis (Kitaev et al., 2022). Herein, we adopt a simple and straightforward one without loss of generality.

Specifically, for a hypothesis sentence H_i, and its corresponding evidence $\{Evi_{i1}, Evi_{i2}, \cdots, Evi_{iK}\}$, we concatenate them all as the input:

$$[CLS] \ H_i \ [SEP] \ Evi_{i1} \ Evi_{i2} \ \cdots \ Evi_{iK} \ [SEP] \tag{1}$$

Then we leverage transformer-based pre-trained language model, i.e., RoBERTa (Liu et al., 2019), to encode the input sequence. Through the multi-head self-attention mechanism (Vaswani et al., 2017), token-level information interaction among the hypothesis sentence and all its evidences is conducted. Besides, since evidence are concentrated, it is much easier to handle the multiple evidence combination issue for cross-sentence inference.

Then the credibility score is calculated through a Multi Layer Perceptron (MLP) with *sigmoid* activation function:

$$\hat{y}_i = \text{Sigmoid}(\text{MLP}(h_i)) \tag{2}$$

where h_i is the hidden state of the special [CLS] token, and is taken as the inference vector of H_i. Besides, $\hat{y}_i \in [0.0, 1.0]$ is the credibility score of H_i, and a higher score means that the sentence is more likely to be entailed by the premise.

2.4 Credibility Fusion

After reading, the inference vector h_i and credibility score $\hat{y}_i$ of each hypothesis sentence H_i is obtained. Nevertheless, the reading model cannot be trained directly since the detailed entailment label of H_i is not available. To this end, we fuse the sentence-level results to judge the entailment relationship between the documents, and indirectly train the model through document-level entailment label. Besides, the fusion process is also expected to solve the detailed disinformation issue for cross-sentence inference, and expand the interpretability of the framework. Herein, we design three fusion methods for comparison.

● **Credibility Score Minimum Pooling** The logic basis for this method is that if the premise entails the hypothesis, then it will entail all the hypothesis sentences, even the one with the lowest credibility score. By contrast, if the premise does not entail the hypothesis, then it will conflict to at least one hypothesis sentence. This one is expected to be assigned with the lowest credibility score.

Formally, for a pair of documents H and P, the credibility scores of the hypothesis sentences are $\{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_m\}$. Then the credibility score of the sample is calculated as:

$$\hat{y}_{HP} = min(\{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_m\}) \tag{3}$$

For this fusion method, the sample prediction result comes from that of the least credible hypothesis sentence, i.e., with the lowest credibility score. The framework is requested to conduct internal contrast among the hypothesis sentences to decide the least credible one. Thus, it is forced to understand the entailment relationship between each hypothesis sentence and its corresponding evidences although without direct entailment label. During prediction, the credibility score $\hat{y}_i$ is utilized to predict the entailment label of hypothesis sentence H_i. We take this as the **main fusion method**.

● **Inference Vector Minimum Pooling** For this method, we conduct minimum pooling on the inference vector h_i rather than the credibility score $\hat{y}_i$. For the inference vector of the sample h_{HP}, the j-th dimension is:

$$h_{HP}^j = min(\{h_1^j, h_2^j, \cdots, h_m^j\}) \tag{4}$$

Then the credibility score of the sample is:

$$\hat{y}_{HP} = \text{Sigmoid}(\text{MLP}(h_{HP})) \tag{5}$$

Many conventional neural models tend to adopt this fusion method for better performance, but suffer from low interpretability, since the practical meaning of each dimension of the inference vector can hardly be probed.

• **Gaussian Kernel Pooling** To further explore the influence of fusion methods, we adopt Gaussian Kernel Pooling (Xiong et al., 2017; Liu et al., 2020; Sheng et al., 2022). Specifically, we utilize C Gaussian kernels $\{K_j\}_{j=1}^{C}$. For a credibility score $\hat{y}_i$, the output of the j-th kernel is:

$$V_i^j = \exp(-\frac{(\hat{y}_i - \mu_j)^2}{2\sigma_j^2}) \tag{6}$$

where μ_j and σ_j are respectively the mean and width of the j-th kernel. In this way, the score $\hat{y}_i$ is projected to a kernel vector $V_i \in \mathcal{R}^C$. The kernel vector of the sample is:

$$V_{HP} = \frac{1}{m}\sum_{i=1}^{m} V_i \in \mathcal{R}^C \tag{7}$$

And the credibility score of the sample is:

$$\hat{y}_{HP} = \text{Sigmoid}(\text{MLP}(V_{HP})) \tag{8}$$

This is a mean pooling method, which conducts mean pooling on the corresponding kernel vectors, rather than the credibility scores.

During training, the loss function of the sample is set as binary cross entropy loss:

$$\mathcal{L} = y_{HP}log(\hat{y}_{HP})+(1-y_{HP})log(1-\hat{y}_{HP}) \tag{9}$$

where y_{HP} is the sample entailment label, 1 for *entailment* samples while 0 for *not entailment* ones. During prediction, we set a threshold on the sample credibility score $\hat{y}_{HP}$ to obtain the result.

3 Experiment

3.1 Dataset

We conduct our experiments on DOCNLI dataset (Yin et al., 2021), which is a newly proposed and the only large scale dataset in the field. The detailed statistic information is shown in Table 1. The training set comes with balanced label distribution, while the development and test sets come with pretty unbalanced ones.

Set	Entailment	Not_Entailment	Total
train	466,653	475,661	942,314
dev	28,890	205,368	234,258
test	33,128	233,958	267,086

Table 1: Statistic information of DOCNLI dataset.

3.2 Complementary Annotation

For interpretability study, we contribute complementary annotation[6]. The hypothesis sentences may involve cross-sentence inference and request multiple evidences. Besides, they may also correspond to several evidence groups, where each group itself is enough to independently infer the entailment label. Thus the annotation process needs heavy workload and comes with great complexity. Herein, we adopt a *proposal and correction* annotation strategy. Specifically, for each hypothesis sentence, we firstly retrieve candidate evidences through the diverse methods in Section 2.2. Then we manually check, remove repeated or unrelated ones and add missing ones, and combine several evidence groups. Finally, we decide the entailment label according to the evidences.

> **Label**: entailment
>
> **Hypothesis Sentence**: Tony Abbott will withdraw Australia's ambassador to Indonesia.
>
> **Evidence Group 1**: Prime Minister Tony Abbott said Australia will withdraw its ambassador to Indonesia in an unprecedented diplomatic response to the executions of Myuran Sukumaran and Andrew Chan.
>
> **Evidence Group 2**: Mr Prasetyo shrugged off diplomatic backlash from Australia after Prime Minister Tony Abbott slammed the executions as "cruel and unnecessary" and announced he would withdraw Australia's ambassador to Indonesia Paul Grigson.

Figure 3: Annotation example for hypothesis sentence. The hypothesis sentence is annotated as *entailment* with two evidence groups.

Due to the heavy workload and great complexity, we manually annotate 100 longer samples (all over 800 words) randomly selected from the *test* set, which contain more than 350 hypothesis sentences in total. An annotation example is shown in Figure 3. The hypothesis sentence is annotated as *entailment* with two evidence groups. More detailed statistic information is summarized in Appendix A.

3.3 Evaluation Metric

For DOCNLI task, we adopt micro and macro F1 scores, and attach more importance to the latter. Due to the unbalanced label distribution, even majority guess model will obtain high micro F1 score, but pretty low macro F1 score. For sentence-level

---

[6]The annotation is also released.

Model	Entailment			Not_Entailment			Total	
	P	**R**	**F1**	**P**	**R**	**F1**	**Micro F1**	**Macro F1**
Concatenation-Longformer$_{base}$♣	-	-	44.42	-	-	-	-	-
Concatenation-Longformer$_{base}$♠	29.73	80.32	43.39	96.33	73.11	83.13	74.01	63.26
Concatenation-RoBERTa$_{base}$♠	39.73	**91.57**	55.41	**98.54**	80.33	88.51	81.72	71.96
Semantic Match-RoBERTa$_{base}$△	31.60	85.49	46.15	97.29	73.80	83.93	75.25	65.04
ours R^{2}F-RoBERTa$_{base}$	**43.52**	87.04	**58.02**	97.86	**84.01**	**90.41**	**84.38**	**74.22**
Concatenation-RoBERTa$_{large}$♣	45.15	**94.99**	61.21	**99.16**	83.66	90.75	85.06	75.98
Concatenation-RoBERTa$_{large}$♠	44.70	94.74	60.74	99.11	83.40	90.58	84.81	75.66
ours R^{2}F-RoBERTa$_{large}$	**50.61**	88.07	**64.28**	98.11	**87.82**	**92.68**	**87.86**	**78.48**

Table 2: Model performance on the *test* set. ♣ and ♠ denote the *original* and *reproduced* results of the model from Yin et al. (2021). △ denotes the *reproduced* results of the model modified from Zhong et al. (2020).

evaluation, we adopt evidence precision, recall and F1 for retrieval, while micro and macro F1 scores for label prediction. Moreover, inspired by Thorne et al. (2018), we propose a more strict metric *full accuracy*. Herein, evidence recall requests to find at least one complete evidence group, and full accuracy further requests correct label prediction.

3.4 Experiment Setup

Our R^{2}F framework is implemented through Pytorch 1.8.0. We adopt AdamW optimizer, keep a random number seed of 42, set max input length as 256, and set mini batch size as 8 with gradient accumulation step as 4. For base encoder, we choose initial learning rate of 1e-5, while for large encoder, we choose 5e-6. For evidence retrieval, we set *K* as 5. During prediction, we adopt a threshold of 0.5. More setup is shown in Appendix B.

3.5 Baseline

Since DOCNLI is still a new task, we adopt the concatenation model from Yin et al. (2021), and modify the semantic match model from Zhong et al. (2020) for comparison. Please refer to the original papers and Appendix C for detailed information.

4 Results

4.1 Main Results

Main results of our framework on the test set are displayed in Table 2. The framework obtains the best performance with the highest micro and macro F1 scores, with both base and large encoders, indicating its strength. The situation is similar on the development set in Appendix D.

All models show far better performance (higher F1 score) on *not entailment* samples than *entailment* samples, while *entailment* samples still come

with even higher recall. This may be due to the label distribution difference among different sets (in Table 1). Among the baselines, semantic match model conducts coarse-grained information interaction between the hypothesis and premise through document-level vector representations. Thus it shows pretty low performance although it can avoid possible key information missing caused by the truncation of overlong samples. This indicates that fine-grained information interaction is essential for this task. Besides, concatenation model with Longformer encoder, although it is able to handle much longer inputs, shows much lower performance than that with RoBERTa encoder, which truncates the inputs. The simplified multi-head self-attention mechanism in Longformer encoder (Beltagy et al., 2020) seems not competent for fine-grained information interaction in the task.

4.2 Performance vs Sample Length

To examine the ability of our framework on processing long inputs, the performance with varying sample length on the test set is shown in Figure 4. Here, the length of a sample is counted in the number of words. Both models obtain far better performance on shorter samples than longer ones (more than 10% absolute difference on both micro F1 and macro F1 scores). Hence it is still a relatively difficult problem to process longer samples. Moreover, our R^{2}F framework consistently and greatly outperforms the strongest concatenation baseline on samples with varying length. Especially, it shows much higher performance on longer samples. These indicate the framework is able to handle longer ones efficiently through breaking the document-level task into sentence-level task with the retrieval, reading and fusion process. On the de-

velopment set, the tendency is similar, and detailed results are displayed in Appendix D.
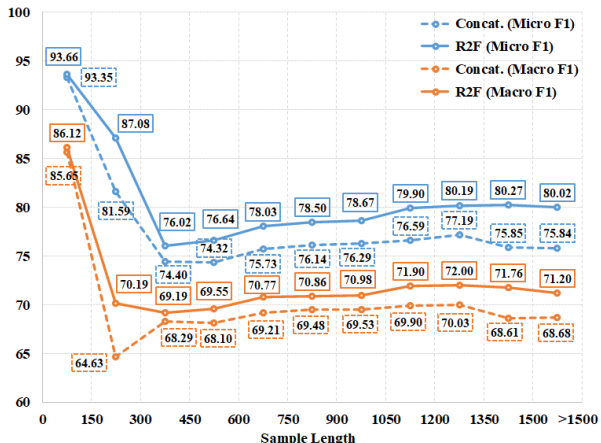

Figure 4: Model performance with varying sample length on the *test* set. Concatenation baseline and R^2F framework with $RoBERTa_{base}$ encoder are compared. The horizontal axis is the sample length in the number of words. The vertical axis is the performance.

4.3 Influence of Evidence Retrieval

To investigate the influence of evidence retrieval, we focus on different retrieval methods, and take those mentioned in Section 2.2 for comparison. Besides, we also adopt a random baseline, which adopts K randomly selected premise sentences as the evidences.

Retrieval	Dev Set		Test Set	
Method	Mi F1	Ma F1	Mi F1	Ma F1
Random	83.27	71.77	82.45	70.98
ROUGE-1	85.58	75.44	84.38	74.22
BM25	85.55	75.36	83.98	73.71
SimCSE*	**86.13**	**75.57**	**85.34**	**74.74**
SimCSE#	84.87	74.33	83.65	73.28

Table 3: DOCNLI performance with different evidence retrieval methods. * and # respectively denote *unsupervised* and *supervised* version, and the same below.

As shown in Table 3, the random baseline can still obtain relatively high performance although it cannot outperform the concatenation baseline. This may be due to the *evidence dependency bias* discussed in Section 4.5. Besides, for shorter samples with only a few premise sentences, evidence retrieval will show less importance. All other retrieval methods can contribute to higher performance than the concatenation baseline, indicating the strength of the framework on the task and its

robustness for diverse retrieval methods. However, supervised SimCSE, although trained on human-annotated NLI benchmarks, shows the lowest performance, which may suffer transfer issue caused by the domain difference between its own training data and DOCNLI dataset.

4.4 Influence of Fusion Method

Performance of different fusion methods are compared in Figure 5. Gaussian kernel pooling tends to incorrectly predicts all samples as *not entailment* and can hardly recognize *entailment* samples. Thus it comes with much higher micro F1 score but far lower macro F1 score under the pretty unbalanced label distribution (in Table 1). This mean pooling fusion method seems not feasible for the task. On both sets, credibility score minimum pooling outperforms all other fusion methods. This fusion method also comes with clear logic basis to expand the interpretability of the framework.
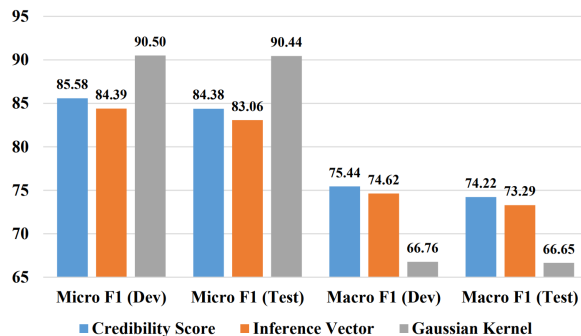

Figure 5: DOCNLI performance with different fusion methods.

4.5 Interpretability Study

To investigate the interpretability of our framework, we conduct sentence-level evaluation on the subset annotated by ourselves. As shown in Table 4, we focus on evidence retrieval and entailment label prediction for hypothesis sentences.

• **Evidence Retrieval** For these longer samples (all over 800 words), the random baseline can hardly obtain the evidence, with extremely low evidence recall. All other retrieval methods can find at least one complete evidence group for most of the hypothesis sentences, with relatively high evidence recall around 85%. However, with pretty low evidence precision and F1 score, all these methods will introduce plenty of uninformative sentences as noise. Thus, the evidence retrieval component may need further improvement.
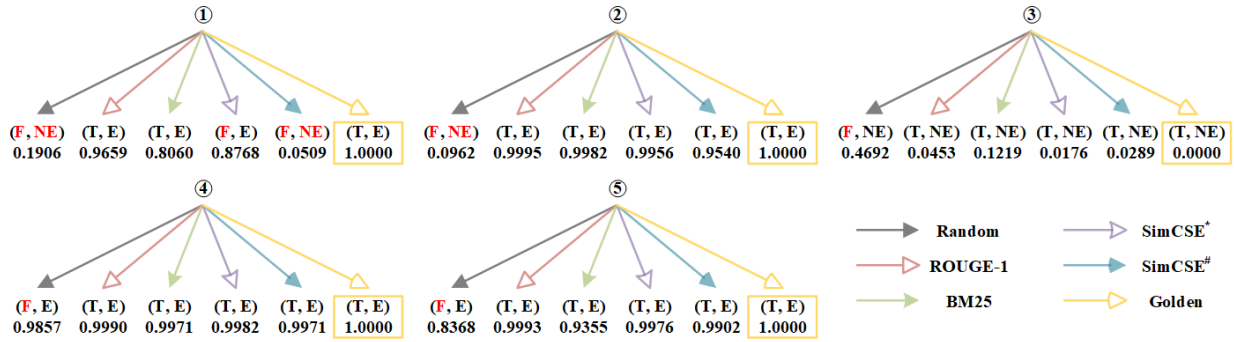
Figure 6: Case study on the sample in Figure 1. For each hypothesis sentence, the elements in the bracket denote whether at least one complete evidence group is obtained, *T* for *True* while *F* for *False*, and entailment label prediction result, *E* for *entailment* while *NE* for *not entailment*. Wrong predictions are in red. The number below is credibility score. The *Golden* branch denotes the groundtruth.

| Retrieval | Evi. Retri. | | | Label Pre. | | FA |
Method	P	R	F1	Mi F1	Ma F1	
Random	6.83	17.21	9.78	66.94	63.32	13.11
ROUGE-1	**31.53**	**87.43**	**46.35**	76.23	73.92	**68.85**
BM25	31.53	87.16	46.31	73.22	70.46	66.12
SimCSE*	30.38	84.97	44.76	**76.50**	**74.69**	67.49
SimCSE#	31.31	84.97	45.76	73.50	70.48	65.03

Table 4: Sentence-level evaluation results. *Evi. Retri.*, *Label Pre.*, and *FA* respectively denote *Evidence Retrieval*, *Label Prediction* and *Full Accuracy*.

• **Entailment Label Prediction** Regardless of poor retrieval performance, the random baseline still shows relatively high performance on entailment label prediction. This is due to *evidence dependency bias*. *Entailment* samples strictly request at least one complete evidence group. However, *not entailment* samples are insensitive to evidence retrieval. With complete evidence group obtained, they are taken as *conflict* to the premise, while with evidence missing, they will be taken as *not mentioned*. Both situations are considered as *not entailment*. This kind of bias will also contribute to the high document-level performance of the random baseline in Table 3. All other retrieval methods obtain much higher performance with the power of evidence. However, it seems that high evidence recall is not promising for more accurate entailment label prediction for hypothesis sentences. This may be also due to evidence dependency bias. Besides, with evidence precision at only around 30%, evidence noise is also an important issue, the influence of which is difficult to estimate. Moreover, the high full accuracy score means for more than 65% hypothesis sentences, our framework can find at least one complete evidence group and correctly predict their entailment label. Therefore, taking that sentence-level annotation is not available during training, our R^2F framework is able to give more interpretable prediction results and help to locate the semantic inconsistency.

4.6 Case Study

To further display the interpretability of our framework, we conduct case study on the sample in Figure 1. As shown in Figure 6, for this long sample (about 667 words in total), the random baseline is pretty weak and cannot find complete evidence group for any hypothesis sentence, although it can obtain correct prediction for the sample. This also demonstrates the importance of evidence discovery for interpretability study. By contrast, all other retrieval methods can obtain complete evidence groups, and contribute to correctly predict sentence-level entailment label for most of the hypothesis sentences. Thus the framework can give more interpretable prediction results, and accurately locate the semantic inconsistency in the third hypothesis sentence. Furthermore, with the two sparse retrieval methods, our framework can even successfully handle all the hypothesis sentences. However, the two dense methods fail to obtain complete evidence group for the first hypothesis sentence. As shown in Figure 1, this one requests two evidences, whose most words are related to the first evidence while only a few words are related to the second one. Moreover, it rephrases the second one with totally different expressions. Therefore, it is difficult to find the second one. This may be the exact situation that leads to the lower evidence recall of dense retrieval methods than sparse ones in Table 4.

5 Related Work

Natural language inference is a fundamental yet important task in natural language processing. For sentence-level NLI, several benchmarks (Bowman et al., 2015; Williams et al., 2018; Wang et al., 2019; Nie et al., 2020) have been proposed and they have been attracting research attention. Moreover, Koreeda and Manning (2021) propose Contract NLI targeting the legal and business domain, which is a small scale benchmark. The premises are long contract documents while the hypotheses are actually single sentences. Besides, Tian et al. (2022) suggest to debias natural language inference and understanding models through causal intervention and counterfactual reasoning. Lin et al. (2022) adopt commonsense inference to enhance future event generation. For document-level NLI, Yin et al. (2021) propose the first large scale benchmark DOCNLI based on a set of early benchmarks. Current models for the task still largely follow sentence-level NLI. Differently, in this paper, we emphasize the importance of evidence discovery and aim at a general solution for the task.

6 Conclusion

In this paper, we propose R^2F framework as a general solution for DOCNLI task and contribute complementary annotation on DOCNLI dataset. Our experimental results show that our framework can obtain state-of-the-art performance. Besides, the framework is robust for diverse retrieval methods, and consistently obtains higher performance on samples with varying length, especially longer samples. Moreover, the framework can give more interpretable prediction results and help to locate the semantic inconsistency. In the future, we will explore an end-to-end framework for the task.

Limitations

The main limitation of our R^2F framework is that the framework is a *pipeline* one rather than an *end-to-end* one. For efficiency, the evidence retrieval component is an independent unit, which provides evidence input for the jointly trained reading and fusion components. However, the evidence retrieval component in our pipeline framework will not bring additional heavy computation. First, it can be conducted offline efficiently. The sparse retrieval methods only involve item frequency counting, and the dense ones are based on pretrained

sentence embeddings without further fine-tuning. Second, there are many acceleration techniques for retrieval in industrial field.

Furthermore, we will try to improve the evidence retrieval component in the future. On one hand, we will try to utilize document-level entailment label to improve sentence-level evidence retrieval. On the other hand, we will also explore an efficient end-to-end model, which may benefit from reinforcement learning (Williams, 1992; Lei et al., 2016), reparameterization trick (Maddison et al., 2017; Jang et al., 2017), or Expectation-Maximization algorithm (Dempster et al., 1977). However, the evidence dependency bias issue, discussed in Section 4.5, will pose a great challenge. That is, *entailment* samples strictly request complete evidence groups, while *not entailment* samples are insensitive to evidence retrieval.

References

Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Chi Kit Cheung. 2022. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. In *Findings of 2022 the Association for Computational Linguistics*, pages 700–710. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Proceedings of the 2015 Advances in Neural Information Processing System*, pages 1171–1179.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. The Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1657–1668. Association for Computational Linguistics.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 2978–2988. Association for Computational Linguistics.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Cogltx: Applying BERT to long texts. In *Proceedings of the 2020 Annual Conference on Neural Information Processing Systems*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910. Association for Computational Linguistics.

Kung Hsiang Huang, Kathleen R. McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2022. Faking fake news for real fake news detection: Propaganda-loaded training data generation. *CoRR*, abs/2203.05386.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations*. OpenReview.net.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks V. S. Lakshmanan. 2022. Automatic detection of entity-manipulated text using factual knowledge. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 86–93. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of 5th International Conference on Learning Representations*. OpenReview.net.

Nikita Kitaev, Thomas Lu, and Dan Klein. 2022. Learned incremental representations for parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3086–3095. Association for Computational Linguistics.

Yuta Koreeda and Christopher D. Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. In *Findings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1907–1919. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117. The Association for Computational Linguistics.

Li Lin, Yixin Cao, Lifu Huang, Shuang Li, Xuming Hu, Lijie Wen, and Jianmin Wang. 2022. What makes the story forward?: Inferring commonsense explanations as prompts for future event generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1098–1109. ACM.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *CoRR*, abs/1907.11692.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351. Association for Computational Linguistics.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *5th International Conference on Learning Representations*. OpenReview.net.

Ziming Mao, Chen Henry Wu, Ansong Ni, Yusen Zhang, Rui Zhang, Tao Yu, Budhaditya Deb, Chenguang Zhu, Ahmed Hassan Awadallah, and Dragomir R. Radev. 2022. Dyle: Dynamic latent extraction for abstractive long-input summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 1687–1698. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901. Association for Computational Linguistics.

Hyunji Hayley Park, Yogarshi Vyas, and Kashif Shah. 2022. Efficient classification of long documents using transformers. In *Proceedings of the 60th Annual*

Meeting of the Association for Computational Linguistics, pages 702–709. Association for Computational Linguistics.

Evan Sandhaus. 2008. The new york times annotated corpus. In *LDC corpora*. Linguistic Data Consortium.

Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom out and observe: News environment perception for fake news detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 4543–4556. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies*, pages 809–819. Association for Computational Linguistics.

Bing Tian, Yixin Cao, Yong Zhang, and Chunxiao Xing. 2022. Debiasing nlu models via causal intervention and counterfactual reasoning. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, pages 11376–11384. AAAI Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 2017 Advances in Neural Information Processing Systems*, pages 5998–6008.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. In *Proceedings of 6th International Conference on Learning Representations*. OpenReview.net.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the7th International Conference on Learning Representations*. OpenReview.net.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-adapter: Infusing knowledge into pre-trained models with adapters. In *Findings of the 2021 Association for Computational Linguistics*, pages 1405–1418. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*, pages

1112–1122. Association for Computational Linguistics.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256.

David H. Wolpert and William G. Macready. 1997. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Hi-transformer: Hierarchical interactive transformer for efficient and effective long document modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 848–853. Association for Computational Linguistics.

Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 55–64. ACM.

Wenpeng Yin, Dragomir R. Radev, and Caiming Xiong. 2021. Docnli: A large-scale dataset for document-level natural language inference. In *Findings of the 2021 Association for Computational Linguistic*, pages 4913–4922. Association for Computational Linguistics.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Hassan Awadallah, Dragomir R. Radev, and Rui Zhang. 2022. $summ^n$: A multi-stage summarization framework for long input dialogues and documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 1592–1604. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 1441–1451. Association for Computational Linguistics.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208. Association for Computational Linguistics.

A Statistic Information of Complementary Annotation

Due to heavy workload and great complexity, we manually annotate 100 longer samples (all over 800 words) randomly selected from the *test* set, which contain more than 350 hypothesis sentences in total. Among the samples, 84% are annotated as

3132

Model	Entailment			Not_Entailment			Total	
	P	**R**	**F1**	**P**	**R**	**F1**	**Micro F1**	**Macro F1**
Concatenation-Longformer$_{base}$♣	-	-	46.18	-	-	-	-	-
Concatenation-Longformer$_{base}$♠	31.09	81.33	44.99	96.60	74.64	84.22	75.47	64.60
Concatenation-RoBERTa$_{base}$♠	42.27	**90.82**	57.69	**98.45**	82.55	89.81	83.57	73.75
Semantic Match-RoBERTa$_{base}$△	31.01	84.32	45.34	97.09	73.61	83.74	74.93	64.54
ours R^{2}F-RoBERTa$_{base}$	**45.54**	86.43	**59.66**	97.82	**85.46**	**91.22**	**85.58**	**75.44**
Concatenation-RoBERTa$_{large}$♣	47.15	**95.09**	63.04	**99.19**	85.01	91.55	86.25	77.30
Concatenation-RoBERTa$_{large}$♠	47.04	94.82	62.89	99.15	84.99	91.52	86.20	77.21
ours R^{2}F-RoBERTa$_{large}$	**53.70**	87.10	**66.43**	98.01	**89.43**	**93.53**	**89.15**	**79.98**

Table 5: Model performance on the *dev* set. ♣ and ♠ denote the *original* and *reproduced* results of the model from Yin et al. (2021). △ denotes the *reproduced* results of the model modified from Zhong et al. (2020).

not entailment, while 16% are annotated as *entailment*. However, among the hypothesis sentences, about 43% are annotated as *not entailment*, while about 57% are annotated as *entailment*. The great label distribution difference is natural. For *entailment* samples, all hypothesis sentences are entailed by the premise. However, *not entailment* samples can still contain entailed hypothesis sentences. Besides, about 47% hypothesis sentences involve cross-sentence inference and request multiple evidence sentences. Moreover, about 17% hypothesis sentences are annotated with more than one evidence groups. These also indicate the great complexity of fine-grained sentence-level annotation.

B Detailed Experiment Setup

Our R^{2}F framework is implemented through Pytorch 1.8.0 and hugging face transformers[7]. All experiments are conducted on a computation node with Nvidia 40G A100 GPUs. For evidence retrieval, we set K as 5 to remain top 5 sentences as evidences during retrieval. We adopt RoBERTa (Liu et al., 2019) as encoder, including base and large version. For all experiments, we adopt AdamW optimizer, keep a random number seed of 42, set max input length as 256, and set mini batch size as 8 with gradient accumulation step as 4. For base encoder, we choose initial learning rate as 1e-5, while for large encoder, we choose initial learning rate as 5e-6. We train 5 epochs, evaluate each 3750 steps, and choose the model parameters with the highest performance on the development set. For Gaussian Kernel Pooling, we keep 11 kernels with the same width of 0.01. However, the mean values of the kernels come from a uniform

---

[7]https://huggingface.co/

distribution within the interval of [0.0, 1.0]. During prediction, we adopt a threshold of 0.5.

C Detailed Baseline

Since DOCNLI is still a new task, we adopt the concatenation model from Yin et al. (2021), and modify the semantic match model from Zhong et al. (2020) for comparison.

• **Concatenation** We concatenate the hypothesis and the premise documents into a sequence as input. Overlong samples will be truncated to max input length. Then we adopt the hidden state of the special [CLS] token as the sample representation for binary classification.

• **Semantic Match** We respectively encode the hypothesis and premise documents to obtain their own document-level vector representation. For documents exceeding the max input length, we split them into chunks with sliding window. Then inspired by Chen et al. (2017), we enhance the vector representations of these two for classification.

D Performance on Development Set

The detailed model performance on the development set are shown in Table 5 and Figure 7. The situations are similar to those on the text set. Our framework obtains the highest performance on the development set. Furthermore, except for the slight performance decrease on samples no longer than 150 words, it greatly and consistently outperforms the strongest concatenation baseline on samples with varying length, especially on longer samples. These also show the strength of our R^{2}F framework on the task.
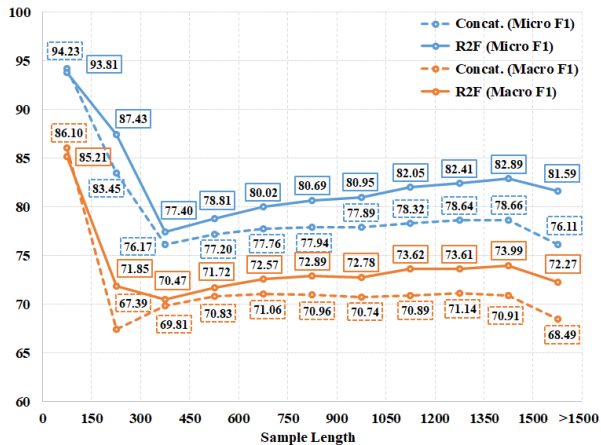
3133

Figure 7: Model performance with varying sample length on the *dev* set. Concatenation baseline and R²F framework with RoBERTa$_{base}$ encoder are compared. The horizontal axis is the sample length in the number of words. The vertical axis is the performance.

K	Dev		Test	
	Mi F1	Ma F1	Mi F1	Ma F1
3	84.79	74.50	83.05	72.65
4	83.68	73.14	81.68	71.30
5	**85.58**	**75.44**	**84.38**	**74.22**
6	84.64	74.50	83.50	73.39
7	84.34	73.95	82.50	72.31

Table 6: DOCNLI performance with varying values of *K*.

E   Influence of *K* Value

To investigate the influence of evidence retrieval, we also pay attention to the value of *K*, which is the hyperparameter about how many sentences to remain during retrieval. As shown in Table 6, the framework is pretty sensitive to the value. Specifically, the highest performance is obtained with *K* as 5. Besides, with *K* as 3, 6 or 7, the framework can also obtain promoted performance than the concatenation baseline. However, with *K* as 4, it shows similar performance with the baseline. As discussed in Section 2.2, *K* is to balance evidence precision and recall. Lower values pursue higher evidence precision, but may lead to evidence missing, while higher values guarantee higher evidence recall, but may introduce too much noise. Besides, the choice of the value is also closely related to data distribution. The high sensitivity to the value indicates that the evidence retrieval process will need further improvement, where a possible way is to utilize the document-level entailment label to improve it.

F   Related Work on Long Document Processing

For long document processing, two common methods are to truncate the input sequence to the maximum length, or cut into several independent segments with sliding window. Dai et al. (2019) propose segment-level recurrence mechanism for information interaction among segments. Beltagy et al. (2020) simplify the self-attention mechanism (Vaswani et al., 2017) to reduce memory overhead to handle longer input sequence. Moreover, Ding et al. (2020) suggest to introduce two independent iteratively trained models, respectively for neural evidence retrieval and semantic inference. Wu et al. (2021) propose hierarchical interactive transformer structure with stacked transformers respectively for sentence and document encoding. However, a recent study (Park et al., 2022) on text classification task indicates that none of these models will consistently outperform other models across datasets. The situation is similar to the famous *No Free Lunch* principle (Wolpert and Macready, 1997). Therefore, long document processing may need further exploration.